ELSEVIER

SciVerse ScienceDirect

Physics Procedia 24 (2012) 1514 – 1519

Physics Procedia

2012 International Conference on Applied Physics and Industrial Engineering

Research On QAR Data Mining Method Based On Improved Association Rule

Qiao yongwei[1],Yang Hui[2], Dong Tingjian[1]

[1] *Engineering & Technical Training Center, Civil Aviation University of China, Tianjin,China,300300*
[2]*College of Computer Science & Technology, Civil Aviation University of China Tianjin,China,300300*

Abstract

The apriori scans database many times and generates large candidate itemsets so that its I/O performance and efficiency seriously affect the universal application and promotion. In order to solve the problem, the paper proposed a method by seting a unique number and recording the location of each itemset. It scans the database once and never generates the huge candidate itemsets and has been applied to QAR data. Experiments show that the proposed algorithm is capable of discovering meaningful and useful association rules in an effective manner, speeding up less execution time.

*Keywords:*Association Rule, Apriori, Frequent Itemsets, Execution Time

1. Introduction

QAR(Quick Access Recorder) is a speed storage devices in the aircraft recording system. It is a scientific and effective technical means to ensure flight safety,operational effciency. Actual engineering practice shows that there have been signs in the QAR before some major accidents occurred. However, the crew is difficult to monitor small changes in parameters because of large volumes of data and many parameters involved, so QAR analysis can not be timely and effective, leading to failure in the subsequent flight to gradually deteriorate, eventually leading to serious consequences. There are of great significnce in the flight technical inspection,security assessment, security incident investigation and aircraft maintenance by using QAR data analysis and mining.

Currently, Western countries on the QAR data research focuses on fault prediction and diagnosis,such as the famous flight operations quality assurance (FOQA) program, but domestic research is still in its infancy[1,2,3]. Association rules is a important method in the data mining field. The basic task in mining for association rules is to determine the correlation between items belonging to a transactional database.

 doi:10.1016/j.phpro.2012.02.224

There has a very important significance on the aircraft fault diagnosis and preventive maintenance through analyzing the QAR data
parameters trends and interrelated relationship using association rules.The apriori is the most important method of association rule mining, but its popularity and promotion have been severely affected because of its poor I/O performance and geometric growth in the number of frequent itemsets. Scholars have obtained relatively good results by carrying out the corresponding improvements from the weighted, fuzzy and other aspects[4-7]. In this paper, we propose a method to record the location of item sets in the data set and simplify the form
of itemsets generation so that it only needs to scan the database once and to reduce the generation of large itemsets. These have increased the efficiency of the algorithm and improve its performance.

2. The Classical Apriori Algorithm

2.1 Association Rules

Given a set of items $I = \{i_1, i_2, ..., i_m\}$ and a database of transactions $D = \{t_1, t_2, ..., t_n\}$ where $t_i = \{I_{i1}, I_{i2}, ..., I_{ip}\}$, $p < m$ and $I_{ij} \in I$,if $X \subseteq I$ with K = |X| is called a k-itemset or simply an itemset. Let a database D be a multi-set of subsets of I as shown. Each $T \in D$ supports an itemset $X \subseteq I$ if $X \subseteq T$ holds. An association rule is an expression $X \to Y$,where X, Y are item sets and $X \bigcap Y = \Phi$ holds. Number of transactions T supporting an item X w.r.t D is alled support of X , $Supp(X) = \left|T \in D \middle| X \subseteq T\right| \Big/ \left|D\right|$. X is a *frequent itemset* f *supp(X)* ≥ σ, where σ (0 ≤ σ ≤ 1) is a predefined inimum support threshold (*minSup*).The strength or onfidence (c) for an association rule $X \to Y$ is the ratio of the number of transactions that contain $X \bigcup Y$ to the number of transactions that contain X , Conf $Conf(X \to Y) = Supp(X \bigcup Y) / Supp(X)$.

2.2 The Classical Apriori Algorithm

Apriori algorithm is the most famous method for association rule mining proposed by Agrawal in 1993. It has two main tasks. First, it can generate the frequent itemsets from the database; second, obtain association rules from the frequent itemsets generated. In fact, the first sub-question is the core in the whole process. When you find all the frequent itemsets, the corresponding association rules would be easy to generate. Maximal frequent itemsets generated as follows.

(1) scan the transaction database D , get a collection of 1-frequent itemsets L_1 .

(2) In general, the assumption L_{K-1} has been generated and now we use it to generate L_k , L_{k-1} connected with itself(item sets and other items in the set of L_{k-1} connected to each other), then we can get the $k-$ candidate itemsets C_k .

(3)Pruning C_k , remove all $(k-1)$ –subsets that not been included in the L_{k-1} itemsets from C_k

(4)Scanning the transaction database D and C_k , we can delete item sets that the count is less than the minimum support in order to get the $k-$ frequent itemsets L_k .

(5)Repeat(2) to (4)until L_k is empty, L_k is the maximal frequent itemsets.

It can be seen that the efficiency is low and the process is complexity through the course of the classical apriori algorithm,mainly due to (1) scaning database many times in the process of generating frequent itemsets, if the database is large, it will seriously affect the algorithm for I/O performance; (2) large candidate itemsets generated affect the efficiency of the algorithm, which will inevitably affect the universal application and promotion of the algorithm.

3. Improved Association Rules Algorithm

In order to improve the performance of Apriori algorithm and its operational efficiency, the paper set for each transaction T a unique number TID in the database D and record the location of each itemset in the database in the process of generating itemsets. Defines K-item sets: $R_k = \langle X_k, TID(X_k) \rangle$,where $X_k = (I_{i1}, I_{i2}, ..., I_{iq})$, $q < m$ and $I_{ij} \in I$, $TIDS(X_{\mathrm{k}})$ is a collection of numbers TID which X_{k} is contained in transaction T in the database. That is to say $TIDS(X_k) = \{TID : X_k \in T, \langle TID, T \rangle \in D\}$.

The support of R_{k} can be expressed as support $(\mathrm{R_k}) = |TIDS(\mathrm{X_k})| / |D| = |\{TID : \mathrm{X_k} \in T, \langle TID, T \rangle \in D| / |D|$. The support number of R_{k} is supNum($\mathrm{R_k}$)=support $(\mathrm{R_k}) * |D| = |TIDS(\mathrm{X_k})|$. After this definition, the process of generating $k-$ itemsets through $(k-1)-$ itemsets can besimplified as: (1)the step of connection. Suppose therer are two $(k-1)-$ itemsets, $L_{k-1}(i) = \langle X_{k-1}, TIDS(X_{K-1}) \rangle$ and $L_{K-1}(j) = \langle Y_{K-1}, TIDS(Y_{k-1}) \rangle$,

respectively, if $X_{k-1}[k-2] = Y_{k-1}[k-2]$,then $L_{k-1}(i)$ and $L_{k-1}(j)$

$$L_{k-1}(i) \propto L_{k-1}(j)$$

connections, that is $= \langle X_{k-1} \bigcup Y_{k-1}, TIDS(X_{k-1}) \bigcap TIDS(Y_{k-1}) \rangle$

$$= \langle X_k, TIDS(X_k) \rangle = R_k \in L_k$$

otherwise, not connection. Because the results are either repeated or non-frequent sets even if we connected them, thus the computation is reduced.(2) According to supNum($\mathrm{R_k}$) = $|TIDS(\mathrm{X_k})|$, we calculate the count of $k-$ itemsets without scaning the database in the pruning process. If supNum($\mathrm{R_k}$)≥minSupNum, then $R_k = \langle X_k, TIDS(X_k) \rangle \in L_k$, Otherwise, remove R_k from the collection L_k . These are tobavoid the I/O operations and increase the efficiency of the algorithm.

Our algorithm only scans the databaseonce,therefore reduces the temporary occupation of memory space as well as I/O cost needed for mining and improve the algorithm effciency, especially in the case of large datasets. In addition, it never generates the huge candidate itemsets in the process of generating $k-$ itemsets through $(k-1)-$ itemsets and avoids the geometric growth in the number of item sets, so it improves the calculation performance during generating the frequent itemsets.

Finally, complete content and organizational editing before formatting. Please take note of the following items when proofreading spelling and grammar:

4. Experimental Evaluation

We have extensively studied our algorithm's performance by comparing it with the classical Apriori algorithm and consider the superiority of it. We performed several experiments using a real QAR data set of B-737-800. Itcontains 11908 records and 256unique items. We select the numberical data contained ALT,CAS,CN1TrakVb1,EGT,N1A1,N2A1 and the algorithm was written in Matlab.

We performed two types of experiments based on quality and performance measures.

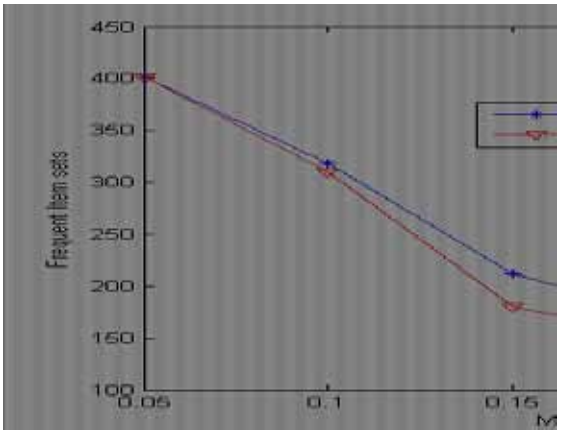

Fig 1 No. of frequent Itemsets with different minsupport

For quality measures, we compared the number of frequent itemsets and the interesting rules generated using two algorithms described above. The results show quite similar behavior of the improved algorithm to classical apriori. As expected the number of frequent itemsets using the improved algorithm are always less than the number of frequent itemsets generated by classical apriori as shown in Figure 1.

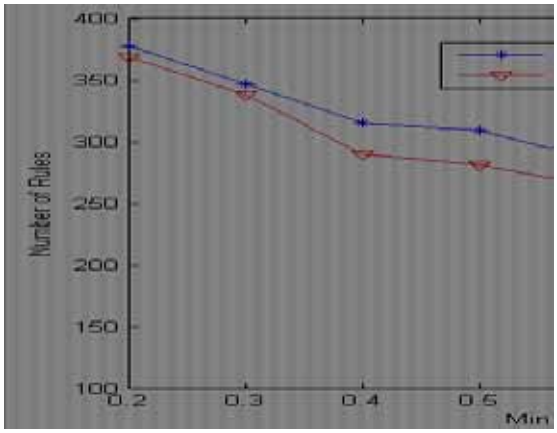

Fig 2 No. of Rules with different confidence

Figure 2 shows the number of interesting rules generated using confidence measures. The classical apriori produces more rules due to the high number of initially generated frequent itemsets .

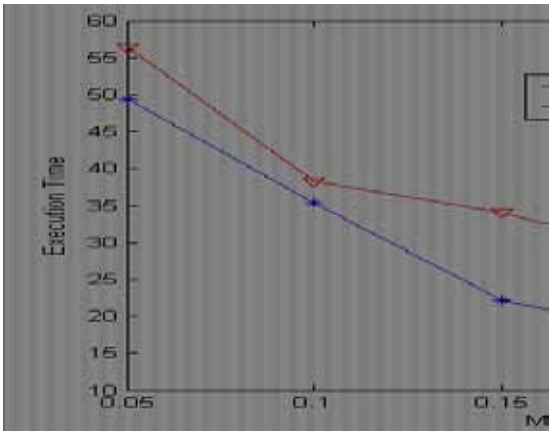

Fig 3 Execution time with different minsupport

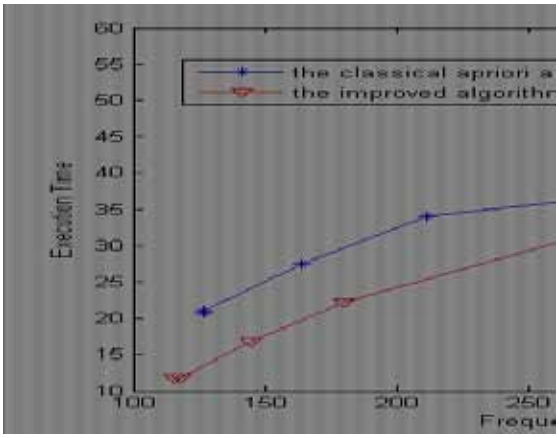

Fig 4 Execution time wirh different frequent itemsets

In the second experiment, we show the scalability of the two algorithms by comparing the execution time with varying the support thresholds. A support threshold from 0.05to 0.3 and confidence 0.6 was used. Figure 3 indicates that the whole execution time of the improved algorithm is better than Apriori. In figure 4 we examine how the performance varies with respect to different number of items while all other things being equal. It indicates that the whole execution time of the improved algorithm is better than Apriori.The execution time increases with the number of items since the more the items, the larger thecandidate itemsets.

5. Conclusion

In this paper, we have presented a improved apriori algorithm for mining association rules by seting a unique number and recording the location of each itemset in the database. It scans the database once and never generates the huge candidate itemsets. The results show that number of frequent itemsets and interesting rules generated using the improved algorithm are less than the classical apriori and spend less execution time. So it is a good way to reduce the I/O cost and improve the algorithm effciency and should be worthy of promotion and application.

6. Acknowledgment

This work has been supported by The project of Civil Aviation Administration of China (MHRD200806).The authors are grateful to the anonymous reviewers for their helpful suggestions and comments..

References

[1] FENG Xing-jie,FENG Xiao-rong,WANG Yan-hua.Application of PCA based on Kernel function in analysis of QAR data.Computer Engineer and Application,2009,45(14):207-209

[2] GENG Hong,JIE Jun.Fuel Flow Regression Model of Aircraft Cruise Based On QAR Data. Aeroengine,2008,34(4):46-50

[3] HUAN Xiu-xia,WANG Hong.Analysis of QAR Data Based On Data Warehouse.Computer Engineering and Design.2008,29(10):2685-2688

[4] Wu Jian, Li Xing-ming. An Effective Mining Algorithm for Weighted Association Rules in Communication Networks. Journal of Computers,2008,3(10):20-26

[5] S.Lotfi, M.H.Sadreddini. Mining Fuzzy Association Rules Using Mutual Information. Proceeding of the International MultiConference of Engineers and Computer Scientists 2009 VolI IMECS2009,March18-20,2009,Hong Kong.

[6] M.Sulaiman Khan, Maybin Muyeba,Frans Coenen. Weighted Association Rule Mining from Binary and Fuzzy Data. Springer-Verlang Berlin Heidelberg 2008,200-212

[7] R. Agrawal and R. Srikant, "Fast algorithms for mining association rules," in Proc. Of the 20th VLDB Conference, Sep. 1994, pp. 478-499.